DOCUMENT RESUME

ED 397 112 TM 025 245

AUTHOR Frey, Sharon L.
TITLE Formulation of an Alternative Method of Determining Levels of Comparison for the Generalized Mantel-Haenszel Using IRT Ability Estimates.
PUB DATE 9 Apr 96
NOTE 41p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Ability; *Comparative Analysis; *Constructed Response; Ethnic Groups; *Item Bias; Item Response Theory; Scores; Simulation; Test Bias; *Test Items
IDENTIFIERS *Graded Response Model; *Mantel Haenszel Procedure; Type I Errors; Type II Errors

ABSTRACT

The Mantel-Haenszel procedure (N. Mantel and W. Haenszel, 1959) and its extension to constructed response items, the Generalized Mantel Haenszel (A. Agresti, 1990), compare performance of subgroups across different score groups to determine differential item functioning (DIF). At each level of comparison, or score group, the subgroups are considered comparable, and the expectation is that the two groups should perform equally well on the item. If there is a significant difference between the subgroups in terms of performance, the item is potentially biased in favor of one subgroup or the other. Traditionally, the levels of comparison have been determined by total raw score. In this paper, an alternative method is proposed for determining the levels of comparison based on an examinee's ability as estimated by the Graded Response Model of F. Samejima (1969). A comparison between the methods using simulated data demonstrates: (1) the high percent of agreement between them; (2) the similar patterns and magnitudes of the overall hit rates and the hit rates for adjusted items; (3) similar patterns and magnitudes for the Type I and Type II error rates; and (4) agreement between the types of items detected by item type and amount of adjustment. (Contains 6 figures, 10 tables, and 7 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
SHARON L. FREY
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 397 112

Formulation of an Alternative Method of Determining Levels of Comparison for the Generalized Mantel-Haenszel using IRT Ability Estimates

Sharon L. Frey

The Riverside Publishing Company

April 9, 1996

Paper presented at the 1996 Annual Meeting of the National Council on Measurement in Education, New York, NY.

TM025245

BEST COPY AVAILABLE

Abstract

The Mantel-Haenszel (Mantel and Haenszel, 1959) and its extension to constructed response items, the Generalized Mantel-Haenszel (Agresti,1990), compare performance of subgroups across different score groups to determine differential item functioning (DIF). At each level of comparison, or score group, the subgroups are considered comparable, and the expectation is that the two groups should perform equally well on the item. If there is a significant difference between the subgroups in terms of performance, the item is potentially biased in favor of one subgroup or the other. Traditionally the levels of comparison have been determined by total raw score. In this paper we propose an alternative method of determining the levels of comparison based on an examinee's ability level as estimated by the Samejima's (1969) Graded Response Model. A comparison between the methods is made to determine the effectiveness of this alternative method.

Formulation of an Alternative Method of Determining Levels of Comparison

for the Generalized Mantel-Haenszel using IRT Ability Estimates

The Mantel-Haenszel (Mantel & Haenszel, 1959) and its extension to constructed response items, the Generalized Mantel-Haenszel (GMH) (Agresti, 1990), compare performance of definable subgroups (reference and focal) across different score groups to determine differential item functioning (DIF). According to Raju, Bode, and Larsen (1989), the optimal number of score groups seems to be a function of the range of scores used to form the score groups. Traditionally the score groups, or the levels of comparison, are determined by using the examinee's total raw score. There could be as few as two score groups or as many score groups as the maximum number of points on the test. At each of these comparison levels, the reference and focal groups are considered comparable. The expectation is that examinees at a particular comparison level, regardless of group membership should perform equally well. That is the pattern of responses should be similar for both subgroups. If there is a significant difference between the reference and focal in terms of performance, the item is potentially biased in favor of one subgroup or the other.

The ways in which examinees can be grouped and compared, however, is not limited to the total raw score received on the test. Using Samejima's (1969) Graded Response model, an indication of the examinee's ability (θ, or theta) can be estimated. This ability estimate not only takes into account the number of items on which the examinee performed well, but also the type of items (i.e., pattern scoring). Did the examinee perform favorably on only the very difficult items? Only the most discriminating items? Taking into consideration that items vary in terms of difficulty and discrimination, implies that there is not always a one-to-one correspondence between raw score and ability estimate. Examinees with the same ability may receive different total test scores. In other words, examinees obtaining the same raw score may vary in terms of their estimated abilities (thetas). According to Lord (1980), the test characteristic curve, the regression line of raw score on ability, specifies the distortion in the ability scale when using the total raw score as a measure of ability.

> Over ability ranges where the test characteristic curve is relatively steep, score differences are exaggerated compared to ability differences. Over ranges where the test characteristic curve is relatively flat, score differences are compressed compared to ability differences. Since number-right scores are integers, compression of a wide range of ability into one or two discrete score values necessarily results in inaccurate measurement. (Lord, 1980), p. 49.

Insert Figure 1 About Here

Figure 1 is a graphical illustration of the possible distortion of using raw score as an ability estimate. Range A, the steepest part of the curve, represents 0.3 difference in ability for a difference of five raw score points. Range B, however, shows a one point difference in ability compressed into one raw score point (raw score = 15) due to the flatness of the curve.

If there is not a one-to-one correspondence between total raw score and ability, and if theta is a better estimate of ability, then is it reasonable to use raw score to identify and compare subgroups when attempting to identify DIF? Could this lead to a mismatch in terms of the comparability of the two subgroups under consideration? Holland and Thayer (1986) explain that measuring the difference in performance on an item between unmatched reference and focal groups is a measure of *impact* rather than differential item performance. The purpose of this study is to compare the results of two methods of determining the levels of comparison for the Generalized Mantel-Haenszel procedure (GMH), one using total raw score and the other using ability parameters. Using simulated data and manipulating certain items to be more difficult for the focal group than the reference group, comparisons between the methods will be made in terms of hit rates, Type I error rates, and Type II error rates.

Method

Sample Design.

The intention of the data simulation design was to assess which method of determining the levels of comparison, ability (theta) or total raw score, was more powerful at detecting simulated bias. It was necessary only to form one reference group (sample size = 500) with an underlying ability distribution that was approximately normal with mean of 0.0 and standard deviation of 1.0. Three underlying ability distributions were generated for

the focal group. Each was approximately normal with standard deviation of 1.0. The means of the distributions were centered at three distinct locations, (a) 0.0 [N(0,1)], (b) -0.50 [N(-.5,1], and (c) -1.0 [N(-1,1)], for four separate sample sizes (50, 100, 250, and 500). Item responses were simulated for each focal group size and focal group location combination for two conditions, control (0 items adjusted) and experimental (9 items adjusted). In the control condition, the discrimination and difficulty item parameters employed for the simulation were identical to the reference group. For the experimental condition, adjustments to the difficulty parameters made the item more difficult for the focal group than for the reference group. Each of the 24 simulated datasets for the focal group [4 (size) x 3 (location) x 2 condition)] were combined with the same simulated dataset for the reference group [size = 500, N(0,1)] to form the datasets for analysis. Table 1 gives the sample design.

Insert Table 1 About Here

Item Design.

In order to yield simulated data that reflected real constructed response items, one of the field test forms of the 1994 Ohio Grade 4 Ohio Proficiency Exam was chosen at random. Each of the Ohio field test forms consisted of a combination of dichotomous (DI), short answer (SA) and extended-response (EX) items, which were worth 1, 2, and 4 points, respectively. The length of the chosen test form was 35 items and consisted of 26 dichotomous, 6 short-answer, and 3 extended-response items. Discrimination and difficulty parameters were generated for this form using the Graded Response Model. These parameters served as a basis for the data simulation for the reference group and the focal group control conditions (0 items adjusted). Displayed in Table 2 are the discrimination (*a* parameters) and difficulty parameters (*b* parameters) for both the control and experimental conditions.

Insert Table 2 About Here

(Note: for the Graded Response Model, the number of difficulty parameters is equal to one less than the total number of possible score categories. Therefore, the number of *b* parameters will vary for each item response type: dichotomous, short answer, and extended response will have 1, 2, and 4 *b* parameters, respectively.) For the

experimental condition, nine items were selected (3 DI, 3 SA, and 3 EX), to be adjusted in difficulty for the focal group. For each item response type, (DI, SA, and EX), the three items selected were adjusted .25, .50, and .75 on the ability scale. In other words, each of the nine items represented a specific combination of item response type and magnitude of adjusted difficulty as shown by the experimental design in Table 3. The manipulated DIF was uniform, that is each *b* parameter for a given item was adjusted by the same magnitude.

Insert Table 3 About Here

Data Simulation.

All item responses were simulated using the Samejima's (1969) Graded Response Model. For ordered responses $x = k$, $k = 1, 2, \ldots m$; where m represents the highest score category, the equation for determining the probability that an item response will be in category k for a given θ, may be written as:

$$P(x = k) = \frac{1}{1 + \exp[-a(\theta - b_{k-1})]} - \frac{1}{1 + \exp[-a(\theta - b_k)]}$$

$$= P^*(k) - P^*(k + 1)$$

In this equation, a is the discrimination parameter, b_k is the b parameter for the kth category, b_{k-1} is the b parameter for the k-1 category, θ is the ability estimate, $P^*(k)$ represents the first term (probability of receiving a score is in category k or higher) and $P^*(k + 1)$ represents the second term (probability of receiving a score in category $(k + 1)$ or higher). For completeness of definition, $P^*(1) = 1$ and $P^*(m + 1) = 0$. For this study, items could be 2-category responses (dichotomous 0-1), 3-category responses (short-answer 0-1-2), or 5-category responses (extended- response 0-1-2-3-4).

For each item, the a and b parameters along with the examinees "underlying" ability estimate (theta, θ) were used to determine the probability of receiving a response in category k or higher ($P^*(k)$ for a given θ). A random number between 0 and 1 drawn from a rectangular distribution was generated and compared to the $P^*(k)$,

starting with the *m*th category. If P*(*m*) was greater than or equal to the random number, the item response simulated was set equal to the score value of the *m*th category. If P*(*m*) was less than the random number, the random number was then compared to P*(*m*-1). If P*(*m*-1) was greater than or equal to the random number, the item response simulated was set equal to the value of the (*m*-1) category. This process continued until an item response was generated. For example, given certain *a* and *b* parameters for a 5-category item and a given theta ($\theta = -.60$), the probability of receiving a score of 4 is P*(5) = .329, probability of receiving a score of 3 or more is P*(4) = .503, 2 or more, P*(3) = .645, 1 or more, P*(2) = .776, and the probability of receiving a score of 0 or more is equal to 1, P*(1) = 1. If the random number generated was equivalent to .575, then the determination of the simulated item response proceeds by first comparing the random number to P*(5) = .329. The random number is greater than .329. Therefore, the comparisons continue with P*(4) = .503, and so on until one of the probabilities (P*(3) = .645) is greater than the random number (.575). In this case the simulated item response would be equal to 2.

Method 1. Each of the 24 datasets (4 (focal size) x 3 (location) x 2 (condition)) was analyzed using the Generalized Mantel-Haenszel with raw scores forming the basis for the levels of comparison. This is the traditional approach and is referred to in this paper as the "raw score" method.

Method 2. Method 2 used the same 24 datasets as in Method 1. These datasets were analyzed a second time using the Generalized Mantel-Haenszel, with the exception that abilities formed the levels of comparison. This method is the alternative method and is referred to in this paper as the "ability" method.

First, the ability estimates for both groups were calibrated together using the computer program, MULTILOG (Thissen, 1986), using the Graded Response Model. MULTILOG is implemented in a two-stage process. In the first stage, the item parameters are estimated based on the total sample. (Note: an increase in the number of iteration cycles from the default of 25 to 40 was necessary to ensure proper convergence.) In the second stage, the item parameters from the first stage were employed in scoring the individual data -- calibration of the ability estimates.

Second, since ability estimates are non-integer values, an appropriate interval had to be chosen to group

BEST COPY AVAILABLE

the ability estimates. The interval had to be chosen in such a manner as to ensure that the interval would be wide enough to include sufficient subgroup sample sizes, yet narrow enough to retain the precision of measurement inherit in the ability scale. Division of the scale into more intervals than there are total score categories is an advantage of the ability scale and a limitation of the raw score method. To determine if increasing the number of intervals (i.e., narrowing the interval width) increases the "precision" of the match of the reference and focal groups in terms of ability, six interval classification methods were devised. For comparison with the raw score method, the first classification method (1) set the number of intervals equal to the same number of intervals in the raw score method. The remaining classification methods increased the number of intervals by (2) + 5 intervals, (3) + 10 intervals, (4) + 15 intervals, (5) + 20 intervals, and (6) + 25 intervals. The interval width for each of the six classification methods was determined by dividing the range of the calibrated ability estimates by the number of total intervals.

Finally, the examinees were assigned to a comparison level based on their estimated ability for each of the six classification methods. The Generalized Mantel-Haenszel procedure was then run six times using the new comparison level groupings. In all cases, for both the raw score and ability methods, items were considered exhibiting DIF if the GMH chi-square statistic was significant at the .01 level.

Figures 2A and 2B illustrate the differences between the groups of examinees compared using the two methods described above. Figure 2A depicts the traditional approach of classifying score groups which ignores the ability dimension and simply defines the levels of comparison based on the raw score (Method 1). Figure 2B illustrates the approach outlined in Method 2 which determines the levels of comparison based exclusively on ability estimates.

Insert Figures 2A and 2B About Here

The items identified as containing DIF by both methods will then be compared by focusing on the following:

1.) Did the two methods agree in terms of the items identified?

2.) Were the items identified with the same statistical significance? Or was one method more powerful than the other?

3.) Was either method more sensitive in identifying the differential functioning of the manipulated items? That is, which method has a better hit rate in terms of correctly identifying those items that were manipulated to have DIF and correctly rejecting items those that were not?

4.) How did the methods agree in terms of identifying the type of items (dichotomous, short-answer, or extended-response) manipulated? Or in terms of identifying the manipulated items based on the magnitude of the adjusted difficulty for the focal group? In other words, is either method more sensitive in detecting items that vary only slightly in difficulty from the reference group?

5.) How did the methods compare in terms of incorrectly detecting the non-DIF items as possessing DIF?

6.) Does varying the number of score groups (levels of comparison) for the ability method affect the detection of the manipulated items?

Results

The percent of agreement between all six interval classification methods with the raw score method for both the control and experimental conditions are shown in Table 4. This table shows the percent of times the methods agree in terms of the direction of significance. In other words, the percentage of times that either the GMH chi-square statistics for both methods were less than or equal to .01 or both were greater than .01, regardless of whether or not the item was actually manipulated for DIF. For the control conditions, the agreement rates between the six interval classification methods and the raw score method ranged from 94% to 100% agreement with most conditions showing 100% agreement. In the experimental conditions, the agreement rates tended to be slightly lower than in the control conditions, and ranged from 86% to 100%, with the majority near 94%.

Insert Table 4 About Here

Insert Table 5 and Figure 3 About Here

Taking into consideration whether or not an item was manipulated to possess DIF, the hit rates (correctly identifying the items that were manipulated and correctly rejecting the items that were not) are shown is Table 5 and graphically represented in Figure 3 (experimental conditions only). Across all control conditions, except for one of the six interval classifications at focal sample size 100 N(0,1) - classification #3, the hit rates for the ability method were greater than or equal to the hit rates for the raw score method. The hit rates for the control condition ranged from 94.3% to 100%, with the majority equal to 100%. For the experimental condition, the hit rates tended to be lower and ranged from 74.3% to 88.6%. In the majority of experimental conditions across interval classification methods, the hit rates for the ability method were greater than or equal to the hit rates for the raw score method. The exceptions to this trend occurred at focal size 100 N(0,1) - classification #3, focal size 250 N(-1,1) - classifications #3 and #5, focal size 500 N(0,1) - classifications #2, #4, #5, and #6, and focal size 500 N(-1, 1) - classifications #1 and #6. There did not appear to be a clear pattern of increased hit rates for smaller ability-intervals (i.e., proceeding from the widest (#1) to the narrowest (#6) interval). Within the focal group location, N(0,1), the hit rates tended to be slightly smaller for larger focal group sizes than for smaller focal group sizes. As the mean underlying ability between the reference group and focal group increased, moving from focal condition N(0,1) to N(-1, 1), the reverse seemed to be true. The hit rates tended to be slightly higher for the larger focal group sizes.

Insert Table 6 and Figure 4 About Here

The error rates for the misclassification of non-DIF items (i.e., not manipulated) as having DIF (Type I error) for both the control and experimental conditions are shown in Table 6 and graphically represented for the experimental condition in Figure 4. The Type I error in the control conditions range from 0.0% to 5.7% with the majority at 0.0%. In only one case is the Type I error for the ability method (2.9%) greater than the raw score method (0.0%). This occurs at focal group size 100 N(0,1) - classification #3. For the experimental condition, the Type I errors range from 0.0% to 14.3%. The zero value Type I error is most frequently occurring in the focal group conditions of sample size 50, although a few zero Type I errors occur for the ability method for larger sample sizes (100 and 250). Examining the table of Type I error rates for the experimental condition shows a

tendency for higher Type I error rates with increases in sample size. This trend seems to hold for all three focal group location conditions (N(0,1), N(-.5,1) and N(-1,1)), although it is more dramatic for the condition N(0,1) and less so for N(-1,1). The largest Type I error rates occur for the sample size 500 N(0,1). In almost all conditions, the Type I error rates for the ability method were less than or equal to the Type I error rates for the raw score method. The only exceptions occurred at sample size 50 N(-.5,1) - classification #5, sample size 500 N(0,1) - classifications #2, #4, #5, #6, and sample size 500 N(-1,1) - classification #1, #4, and #6. There does not seem to a be any pattern of decreased error rates for smaller ability-intervals (i.e., proceeding from the widest (#1) to the narrowest (#6) interval).

Insert Table 7 and Figure 5 About Here

The error rates for the misclassification of DIF items (i.e., manipulated) as being non-DIF (Type II error) for experimental conditions are shown in Table 7 and graphically represented in Figure 5. The Type II errors range from 8.6% to 22.9%. In general, the Type II error rates tend to decrease slightly with increasingly larger sample sizes. The trend seems to be hold regardless of the focal group location (N(0,1), N(-.5,1), N(-1,1)). The largest Type II errors occur for smaller focal group sizes, in particular condition N(-1,1), while the smallest Type II errors occur for larger focal group sizes. As the focal group size increases, the Type II error rates are more uniform across focal group conditions (N(0,1), N(-.5,1) and N(-1,1)) than for smaller focal group sizes. In almost all cases, the Type II error rates for the ability method were less than or equal to the Type II error rates for the raw score method. The only exceptions occurred at sample size 50 N(-.1,1) - classifications #4 and #5, sample size 100 N(0,1) - classifications #3, and sample size 250 N(-1,1) - classification #2, #3, and #5. The Type II error rates seem to be, to some extent, much more uniform across interval classification methods than the Type I error rates discussed previously.

Insert Table 8 and Figure 6 About Here

The hit rates based on only the 9 adjusted items for the experimental condition are displayed in Table 8 and graphically represented in Figure 6. The hit rates tended to be higher for larger focal group sizes for all three focal group location conditions (N(0,1), N(-.5,1) and N(-1,1)). The difference in hit rates between small and large focal group sizes appeared to be slightly greater within the focal group location N(-1,1) than for the conditions N(0,1) or N(-.5, 1). The ability method hit rates of the adjusted items were greater than or equal to the raw score method, except for a few instances. These exceptions were at focal group size 100 N(0,1) - classification #3, and at the focal group size 250 N(-1,1), classifications #2, #3, and #5.

Insert Table 9 About Here

Table 9 illustrates which of the 9 items manipulated in the experimental conditions were correctly identified as possessing DIF. The "X" indicates which items were correctly identified for each item response type (DI, SA, and EX) and magnitude of adjustment (.25, .50, and .75) combination. The table indicates that the items that were adjusted by .25, regardless of item type, were rarely detected. In fact, in only two cases was an item adjusted .25 ever detected, one was a dichotomous item in the focal group condition 50 N(-1,1) and the other was an extended response item in the focal group condition 500 N(-1,1). Short-answer items were typically only detected if the items were adjusted by .75, and detection rates increased with larger focal group sample sizes. The only case in which a short-answer item adjusted by .50 was detected occurred for the focal group condition 100 N(-1,1). Dichotomous items were detected more often at the adjustment magnitudes of .50 and .75, especially in the focal group conditions N(0,1) and N(-.5,1) for smaller sample sizes, and in focal group conditions N(0,1), N(-.5,1) and N(-1,1) for larger sample sizes. For extended-response items, only items adjusted by .75 tended to be detected at smaller sample sizes, those adjusted by .50 and .75 were detected at larger sample sizes. It appears that extended-response items may be more likely detected in the conditions N(0,1) and N(-.5,1) than in the condition N(-1,1), with detection rates increasing slightly for N(-1,1) with increasing focal group sample sizes. In comparison to the raw score method, the overall detection capability of the ability classifications seemed to be the same or better than the raw score method, with a few minor exceptions. For the short-answer and extended-response items, there was only one case each in which the ability method did not detect a manipulated item

(adjusted .75) that the raw score method was able to detect. For the short-answer item, this occurred for focal group condition 50 N(-.5,1) - classification #4. The missed detection for the extended-response item occurred for the focal group condition 100 N(0,1) - classification #3. For dichotomous items, there were a few instances in which the raw score method out performed the ability method in terms of identifying items manipulated by .50. This happened for the focal group condition 100 N(-1,1) - for all classifications, and for 250 N(-1,1) - classifications #3 and #6.

It is not only important to examine the detection of manipulated items, but it is also important to investigate the detection of the non-manipulated items (non-DIF items). The percentages of non-DIF items that were falsely detected are shown in Table 10 for both the control and experimental conditions. The values are expressed as percents and are based on the total number of non-DIF items (total = 26). This table also shows the type of items that were misidentified (DI or SA). Note, since there were only three extended-response items for the test form chosen, those extended-response items were used in the experimental condition and all three manipulated for DIF. Therefore, no extended-response items are shown as misidentified as having DIF.

Insert Table 10 About Here

In the control condition, the ability classifications always did the same or better at not misidentifying the non-DIF items. For the experimental conditions, the ability method always did the same or better at not falsely detecting short-answer items. However, for dichotomous items in the experimental conditions, this is not always the case. In particular, the ability method misdetected more non-DIF dichotomous items in the focal group conditions: sample size 50 N(-.5,1) - classification #5, sample size 50 N(-1,1) - classification #2, sample size 500 N(0,1) - classification #2, #4, #5, and #6, and sample size 500 N(-1,1) - classification #1,#2,#3,#4, and #6. The misdetection rates are higher for the sample sizes of 500 than for the smaller sample sizes of 50. Moreover, the misdetection was the highest for 500 N(0,1). The erroneous detection of dichotomous items increases dramatically with sample size for the focal group location N(0,1), the same trend tends to occur for N(-.5,1) and N(-1,1), but with less dramatic increases.

BEST COPY AVAILABLE

Summary and Conclusion

One important result from this study is that overall the results from both the ability and raw score methods were very comparable. This is demonstrated by 1) the high percent of agreement between the methods (averaging 99% in control condition and 94% in the experimental condition), 2) similar patterns and magnitudes of the overall hit rates and the hit rates for only the adjusted items, 3) similar patterns and magnitudes for both the Type I and Type II error rates, and 4) the agreement among the types of items detected by item type and amount of adjustment. In general, it appears that the items identified by both methods were identified with the same statistical significance. Regardless of item type, items that were adjusted by .25 on the ability metric were rarely ever detected by either method, while the items adjusted by .75 were almost always detected, especially with larger focal group sizes. Even in the smallest sample sizes, both methods were sensitive enough to detect manipulated differences of .50 for dichotomous items, notably when the reference and focal group were closer in terms of underlying ability. Only in larger focal group sizes were extended-response items adjusted by .50 ever detected. The ability method tended to erroneously detect more dichotomous non-DIF items as DIF than the raw score method, markedly at larger focal group sizes. However, no significant difference between the methods occurred in terms of misidentifying non-DIF short-answer items. Changing the interval widths for the ability method seemed to have more of an affect on the Type I error than Type II error. Within the same focal group condition, the Type II error appeared to be consistent across interval width classifications, whereas the Type I error was more variable. That is, decreasing the size of the ability interval used for forming the levels of comparison did not improve the detection of the manipulated items. There were some slight differences between the methods, but without further research and more replications of the process, no firm conclusions as to which method is better can be drawn. In fact, if anything, this study proves just how powerful the GMH can be when compared to a method in which the levels of comparison are based on abilities, the two methods are very comparable in terms of results.

References

Agresti, A. (1990). Categorical Data Analysis. New York: Wiley.

Holland, P. W., & Thayer, D. T. (1986, April). *Differential item performance and the Mantel-Haenszel procedure*. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mantel, N., & Haenszel, W. (1959). *Statistical aspects of the analysis of data from retrospective studies of disease*. Journal of the National Cancer Institute, 22, 719-748.

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). *An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance*. Applied Measurement in Education, 2(1), 1-13.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometrika Monograph Supplement, 4, Part 2, Whole #17.

Thissen, D. (1986). MULTILOG User's Manual. Mooresville, IN: Scientific Software.

Table 1. Sample design: characteristics of the focal group by size and location for both the control and experimental conditions.

CONTROL (0 ITEMS ADJUSTED -- FOCAL GROUP)

REFERENCE GROUP: SIZE = 500, N(0,1)

FOCAL GROUP LOCATION		FOCAL GROUP SIZE 50	100	250	500
	(A) N(0.0, 1)				
	(B) N(-0.5, 1)				
	(C) N(-1.0, 1)				

EXPERIMENTAL (9 ITEMS ADJUSTED -- FOCAL GROUP)

REFERENCE GROUP: SIZE = 500, N(0,1)

FOCAL GROUP LOCATION		FOCAL GROUP SIZE 50	100	250	500
	(A) N(0.0, 1)				
	(B) N(-0.5, 1)				
	(C) N(-1.0, 1)				

TABLE 2: Target discrimination (A) and difficulty (B) item parameters for the control and experimental conditions.

			TARGET B ITEM PARAMETERS -- CONTROL				FOCAL GROUP	TARGET B ITEM PARAMETERS -- EXPERIMENTAL			
ITEM	PTS	A	B(1)	B(2)	B(3)	B(4)	Adj.	B(1)	B(2)	B(3)	B(4)
1	1	0.910	-0.620				.25	-0.370			
2	1	0.970	-2.120				.00	-2.120			
3	1	1.860	-1.280				.00	-1.280			
4	1	1.200	-1.250				.00	-1.250			
5	1	1.680	-0.940				.00	-0.940			
6	2	0.790	-4.550	-0.370			.25	-4.300	-0.120		
7	1	1.970	-1.370				.00	-1.370			
8	1	1.280	-1.980				.00	-1.980			
9	4	1.170	-1.660	-1.110	-0.610	0.010	.25	-1.410	-0.860	-0.360	0.260
10	2	1.200	-2.740	-1.380			.50	-2.240	-0.880		
11	1	1.860	-1.630				.00	-1.630			
12	1	1.510	-1.500				.00	-1.500			
13	1	1.110	-0.910				.00	-0.910			
14	1	1.770	-1.450				.00	-1.450			
15	1	1.160	-1.350				.00	-1.350			
16	1	0.410	-0.650				.00	-0.650			
17	1	1.370	-1.330				.00	-1.330			
18	1	2.300	-1.350				.50	-0.850			
19	1	0.850	-0.280				.00	-0.280			
20	2	0.690	-5.000	-0.590			.00	-5.000	-0.590		
21	1	1.720	-1.800				.75	-1.050			
22	1	1.350	-1.280				.00	-1.280			
23	1	0.890	-1.260				.00	-1.260			
24	1	1.590	-1.100				.00	-1.100			
25	2	1.170	-1.810	-0.030			.00	-1.810	-0.030		
26	1	1.210	-1.210				.00	-1.210			
27	4	1.560	-1.700	-0.300	0.920	2.200	.50	-1.200	0.200	1.420	2.700
28	1	0.840	-1.340				.00	-1.340			
29	4	1.520	-1.280	-0.670	0.060	0.990	.75	-0.530	0.080	0.810	1.740
30	1	0.960	-1.610				.00	-1.610			
31	1	1.010	-0.870				.00	-0.870			
32	1	1.280	-1.170				.00	-1.170			
33	1	1.920	-1.150				.00	-1.150			
34	2	0.450	-1.270	2.400			.00	-1.270	2.400		
35	2	1.280	-0.720	0.420			.75	0.030	1.170		

BEST COPY AVAILABLE

Table 3. Experimental condition -- manipulated DIF for the focal group only. Manipulated item design by item type and magnitude of adjustment.

(DIFFICULTY PARAMETERS)

ITEM TYPE	ADJUSTMENT MAGNITUDE .25	.50	.75
DI	1	1	1
SA	1	1	1
EX	1	1	1

9

DI = DICHOTOMOUS (0-1 POINTS)
SA = SHORT ANSWER (0-2 POINTS)
EX = EXTENDED RESPONSE (0-4 POINTS)

Table 4. Percent of agreement with the raw score method in terms of the direction of significance.

REFERENCE	FOCAL GROUP		CONTROL (0 ITEMS ADJUSTED)						EXPERIMENTAL (9 ITEMS ADJUSTED)					
SIZE	SIZE	LOCATION	1	2	3	4	5	6	1	2	3	4	5	6
500	50	N(0.0, 1)	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.97	0.97	0.97	0.97	0.97
500	50	N(-0.5, 1)	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.94	0.94	0.94	0.91	0.94
500	50	N(-1.0, 1)	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.94	0.97	0.94	0.94	0.97
500	100	N(0.0, 1)	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00
500	100	N(-0.5, 1)	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.91	0.97	0.91	0.94	0.91
500	100	N(-1.0, 1)	1.00	1.00	1.00	1.00	1.00	1.00	0.89	0.89	0.94	0.94	0.91	0.91
500	250	N(0.0, 1)	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.91	0.94	0.94	0.97
500	250	N(-0.5, 1)	0.97	0.97	1.00	1.00	1.00	1.00	0.86	0.86	0.91	0.91	0.94	0.91
500	250	N(-1.0, 1)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.97	1.00	0.97	0.91
500	500	N(0.0, 1)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.97	0.97	0.97
500	500	N(-0.5, 1)	0.97	0.97	0.97	0.97	0.97	0.97	0.91	0.91	0.91	0.94	0.91	0.91
500	500	N(-1.0, 1)	1.00	1.00	1.00	1.00	1.00	1.00	0.91	0.94	0.91	0.94	0.97	0.97

Table 5. Overall hit rates for raw score and ability methods.

REFERENCE SIZE	FOCAL GROUP SIZE	FOCAL GROUP LOCATION	CONTROL (0 ITEMS ADJUSTED) 1	2	3	4	5	6	RS	EXPERIMENTAL (9 ITEMS ADJUSTED) 1	2	3	4	5	6	RS
500	50	N(0.0, 1)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.829	0.829	0.829	0.829	0.829	0.829	0.800
500	50	N(-0.5, 1)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.857	0.857	0.857	0.800	0.829	0.857	0.800
500	50	N(-1.0, 1)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.771	0.743	0.771	0.800	0.800	0.771	0.743
500	100	N(0.0, 1)	1.000	1.000	0.971	1.000	1.000	1.000	1.000	0.829	0.829	0.800	0.829	0.829	0.829	0.829
500	100	N(-0.5, 1)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.829	0.800	0.800	0.800	0.829	0.800	0.771
500	100	N(-1.0, 1)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.800	0.800	0.800	0.800	0.829	0.829	0.800
500	250	N(0.0, 1)	1.000	1.000	1.000	1.000	1.000	1.000	0.943	0.829	0.829	0.800	0.829	0.829	0.800	0.771
500	250	N(-0.5, 1)	1.000	1.000	0.971	0.971	0.971	0.971	0.971	0.829	0.829	0.829	0.829	0.857	0.886	0.800
500	250	N(-1.0, 1)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.829	0.829	0.800	0.829	0.800	0.857	0.829
500	500	N(0.0, 1)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.771	0.743	0.771	0.743	0.743	0.743	0.771
500	500	N(-0.5, 1)	1.000	1.000	1.000	1.000	1.000	1.000	0.971	0.800	0.800	0.800	0.771	0.800	0.800	0.771
500	500	N(-1.0, 1)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.800	0.829	0.857	0.829	0.857	0.800	0.829

BEST COPY AVAILABLE

Table 6. Type I error rates for the raw score and ability methods -- control and experimental conditions.

REFERENCE	FOCAL GROUP		CONTROL (0 ITEMS ADJUSTED)							EXPERIMENTAL (9 ITEMS ADJUSTED)						
SIZE	SIZE	LOCATION	1	2	3	4	5	6	RS	1	2	3	4	5	6	RS
500	50	N(0.0, 1)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
500	50	N(-0.5, 1)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.029	0.000	0.000
500	50	N(-1.0, 1)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.029	0.000	0.000	0.000	0.000	0.029
500	100	N(0.0, 1)	0.000	0.000	0.029	0.000	0.000	0.000	0.000	0.029	0.029	0.029	0.029	0.029	0.029	0.029
500	100	N(-0.5, 1)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.029	0.029	0.029	0.000	0.029	0.029
500	100	N(-1.0, 1)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.029	0.029	0.029	0.029	0.000	0.000	0.029
500	250	N(0.0, 1)	0.000	0.000	0.000	0.000	0.000	0.000	0.057	0.057	0.057	0.086	0.057	0.057	0.086	0.114
500	250	N(-0.5, 1)	0.000	0.000	0.029	0.029	0.029	0.029	0.029	0.057	0.057	0.057	0.057	0.029	0.000	0.086
500	250	N(-1.0, 1)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.029	0.000	0.029	0.029	0.029	0.000	0.029
500	500	N(0.0, 1)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.114	0.143	0.114	0.143	0.143	0.143	0.114
500	500	N(-0.5, 1)	0.000	0.000	0.000	0.000	0.000	0.000	0.029	0.086	0.086	0.086	0.114	0.086	0.086	0.114
500	500	N(-1.0, 1)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.086	0.057	0.057	0.086	0.029	0.086	0.057

Table 7. Type II error rates for the raw score and ability methds -- experimental condition.

REFERENCE	FOCAL GROUP		EXPERIMENTAL (9 ITEMS ADJUSTED)						
SIZE	SIZE	LOCATION	1	2	3	4	5	6	RS
500	50	N(0.0, 1)	0.171	0.171	0.171	0.171	0.171	0.171	0.200
500	50	N(-0.5, 1)	0.143	0.143	0.143	0.200	0.143	0.143	0.200
500	50	N(-1.0, 1)	0.229	0.229	0.229	0.200	0.200	0.229	0.229
500	100	N(0.0, 1)	0.143	0.143	0.171	0.143	0.143	0.143	0.143
500	100	N(-0.5, 1)	0.171	0.171	0.171	0.171	0.171	0.171	0.200
500	100	N(-1.0, 1)	0.171	0.171	0.171	0.171	0.171	0.171	0.171
500	250	N(0.0, 1)	0.114	0.114	0.114	0.114	0.114	0.114	0.114
500	250	N(-0.5, 1)	0.114	0.114	0.114	0.114	0.114	0.114	0.114
500	250	N(-1.0, 1)	0.143	0.171	0.171	0.143	0.171	0.143	0.143
500	500	N(0.0, 1)	0.114	0.114	0.114	0.114	0.114	0.114	0.114
500	500	N(-0.5, 1)	0.114	0.114	0.114	0.114	0.114	0.114	0.114
500	500	N(-1.0, 1)	0.114	0.114	0.086	0.086	0.114	0.114	0.114

Table 8. Hit rates for adjusted items only -- raw score and ability methods.

REFERENCE SIZE	FOCAL GROUP		EXPERIMENTAL (9 ITEMS ADJUSTED)						
	SIZE	LOCATION	1	2	3	4	5	6	RS
500	50	N(0.0, 1)	0.333	0.333	0.333	0.333	0.333	0.333	0.222
500	50	N(-0.5, 1)	0.444	0.444	0.444	0.222	0.444	0.444	0.222
500	50	N(-1.0, 1)	0.111	0.111	0.111	0.222	0.222	0.111	0.111
500	100	N(0.0, 1)	0.444	0.444	0.333	0.444	0.444	0.444	0.444
500	100	N(-0.5, 1)	0.333	0.333	0.333	0.333	0.333	0.333	0.222
500	100	N(-1.0, 1)	0.333	0.333	0.333	0.333	0.333	0.333	0.333
500	250	N(0.0, 1)	0.556	0.556	0.556	0.556	0.556	0.556	0.556
500	250	N(-0.5, 1)	0.556	0.556	0.556	0.556	0.556	0.556	0.556
500	250	N(-1.0, 1)	0.444	0.333	0.333	0.444	0.333	0.444	0.444
500	500	N(0.0, 1)	0.556	0.556	0.556	0.556	0.556	0.556	0.556
500	500	N(-0.5, 1)	0.556	0.556	0.556	0.556	0.556	0.556	0.556
500	500	N(-1.0, 1)	0.556	0.556	0.667	0.667	0.556	0.556	0.556

Table 9. Items identified as DIF by item type and magnitude of adjustment.

REFERENCE SIZE	FOCAL GROUP SIZE	FOCAL GROUP LOCATION	METHOD	DICHOTOMOUS .25	DICHOTOMOUS .50	DICHOTOMOUS .75	SHORT ANSWER .25	SHORT ANSWER .50	SHORT ANSWER .75	EXTENDED RESPONSE .25	EXTENDED RESPONSE .50	EXTENDED RESPONSE .75
500	50	N(0.0, 1)	1		X	X						X
			2		X	X						X
			3		X	X						X
			4		X	X						X
			5		X	X						X
			6		X	X						X
			RS		X							
500	50	N(-0.5, 1)	1		X	X			X			X
			2		X	X			X			X
			3		X	X			X			X
			4		X	X						
			5		X	X			X			X
			6		X	X			X			X
			RS		X				X			
500	50	N(-1.0, 1)	1	X								
			2	X								
			3	X						X		
			4	X						X		
			5	X								
			6	X								
			RS	X								

REFERENCE SIZE	FOCAL GROUP SIZE	FOCAL GROUP LOCATION	METHOD	DICHOTOMOUS .25	DICHOTOMOUS .50	DICHOTOMOUS .75	SHORT ANSWER .25	SHORT ANSWER .50	SHORT ANSWER .75	EXTENDED RESPONSE .25	EXTENDED RESPONSE .50	EXTENDED RESPONSE .75
500	100	N(0.0, 1)	1		X	X			X			X
			2		X	X			X			X
			3		X	X			X			
			4		X	X			X			X
			5		X	X			X			X
			6		X	X			X			X
			RS		X	X			X			
500	100	N(-0.5, 1)	1		X	X			X			
			2		X	X			X			
			3		X	X			X			
			4		X	X			X			
			5		X	X			X			
			6		X	X			X			
			RS			X			X			
500	100	N(-1.0, 1)	1				X	X				X
			2				X	X				X
			3				X	X				X
			4				X	X				X
			5				X	X				X
			6					X				X
			RS		X							

BEST COPY AVAILABLE

Table 9. Items identified as DIF by item type and magnitude of adjustment. (Continued)

REFERENCE SIZE	FOCAL GROUP SIZE	FOCAL GROUP LOCATION	METHOD	DICHOTOMOUS .25	DICHOTOMOUS .50	DICHOTOMOUS .75	SHORT ANSWER .25	SHORT ANSWER .50	SHORT ANSWER .75	EXTENDED RESPONSE .25	EXTENDED RESPONSE .50	EXTENDED RESPONSE .75
500	250	N(0.0, 1)	1		X	X			X		X	X
			2		X	X			X		X	X
			3		X	X			X		X	X
			4		X	X			X		X	X
			5		X	X			X		X	X
			6		X	X			X		X	X
			RS		X	X			X		X	X
500	250	N(-0.5, 1)	1		X	X			X		X	X
			2		X	X			X		X	X
			3		X	X			X		X	X
			4		X	X			X		X	X
			5		X	X			X		X	X
			6		X	X			X		X	X
			RS		X	X			X		X	X
500	250	N(-1.0, 1)	1		X	X			X			X
			2			X			X			X
			3			X			X			X
			4		X	X			X			X
			5			X			X			X
			6			X			X		X	X
			RS		X	X			X			X

REFERENCE SIZE	FOCAL GROUP SIZE	FOCAL GROUP LOCATION	METHOD	DICHOTOMOUS .25	DICHOTOMOUS .50	DICHOTOMOUS .75	SHORT ANSWER .25	SHORT ANSWER .50	SHORT ANSWER .75	EXTENDED RESPONSE .25	EXTENDED RESPONSE .50	EXTENDED RESPONSE .75
500	500	N(0.0, 1)	1		X	X			X		X	X
			2		X	X			X		X	X
			3		X	X			X		X	X
			4		X	X			X		X	X
			5		X	X			X		X	X
			6		X	X			X		X	X
			RS		X	X			X		X	X
500	500	N(-0.5, 1)	1		X	X			X		X	X
			2		X	X			X		X	X
			3		X	X			X		X	X
			4		X	X			X		X	X
			5		X	X			X		X	X
			6		X	X			X		X	X
			RS		X	X			X		X	X
500	500	N(-1.0, 1)	1		X	X			X		X	X
			2		X	X			X		X	X
			3		X	X			X	X	X	X
			4		X	X			X	X	X	X
			5		X	X			X		X	X
			6		X	X			X		X	X
			RS		X	X			X		X	X

Table 10. Items misidentified as DIF by item type.
Note: Percents based on the 26 non-adjusted items: DI = 23, SA = 3

CONTROL CONDITION (0 ITEMS ADJUSTED -- FOCAL GROUP)

REFERENCE SIZE	FOCAL SIZE	METHOD	FOCAL GROUP LOCATION N(0, 1) % TOTAL	% DI	% SA	N(-.5, 1) % TOTAL	% DI	% SA	N(-1, 1) % TOTAL	% DI	% SA
500	50	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		RS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	100	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		RS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500	250	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		3	0.00	0.00	0.00	3.85	4.35	0.00	0.00	0.00	0.00
		4	0.00	0.00	0.00	3.85	4.35	0.00	0.00	0.00	0.00
		5	0.00	0.00	0.00	3.85	4.35	0.00	0.00	0.00	0.00
		6	0.00	0.00	0.00	3.85	4.35	0.00	0.00	0.00	0.00
		RS	7.69	4.35	33.33	3.85	4.35	0.00	0.00	0.00	0.00
500	500	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		RS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

EXPERIMENTAL CONDITION (9 ITEMS ADJUSTED -- FOCAL GROUP)

REFERENCE SIZE	FOCAL SIZE	METHOD	FOCAL GROUP LOCATION N(0, 1) % TOTAL	% DI	% SA	N(-.5, 1) % TOTAL	% DI	% SA	N(-1, 1) % TOTAL	% DI	% SA
500	50	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		2	0.00	0.00	0.00	0.00	0.00	0.00	3.85	4.35	0.00
		3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		5	0.00	0.00	0.00	3.85	4.35	0.00	0.00	0.00	0.00
		6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		RS	0.00	0.00	0.00	0.00	0.00	0.00	3.85	0.00	33.33
500	100	1	3.85	4.35	0.00	0.00	0.00	0.00	3.85	4.35	0.00
		2	3.85	4.35	0.00	3.85	4.35	0.00	3.85	4.35	0.00
		3	3.85	4.35	0.00	3.85	4.35	0.00	3.85	4.35	0.00
		4	3.85	4.35	0.00	3.85	4.35	0.00	3.85	4.35	0.00
		5	3.85	4.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		6	3.85	4.35	0.00	3.85	4.35	0.00	0.00	0.00	0.00
		RS	3.85	4.35	0.00	3.85	4.35	0.00	3.85	4.35	0.00
500	250	1	7.69	4.35	33.33	7.69	8.70	0.00	3.85	4.35	0.00
		2	7.69	4.35	33.33	7.69	8.70	0.00	0.00	0.00	0.00
		3	11.54	8.70	33.33	7.69	8.70	0.00	3.85	4.35	0.00
		4	7.69	4.35	33.33	7.69	8.70	0.00	3.85	4.35	0.00
		5	7.69	4.35	33.33	3.85	4.35	0.00	3.85	4.35	0.00
		6	11.54	8.70	33.33	0.00	0.00	0.00	0.00	0.00	0.00
		RS	15.38	13.04	33.33	11.54	13.04	0.00	3.85	4.35	0.00
500	500	1	15.38	17.39	0.00	11.54	13.04	0.00	11.54	13.04	0.00
		2	19.23	21.74	0.00	11.54	13.04	0.00	7.69	8.70	0.00
		3	15.38	17.39	0.00	11.54	13.04	0.00	7.69	8.70	0.00
		4	19.23	21.74	0.00	15.38	13.04	33.33	11.54	8.70	33.33
		5	19.23	21.74	0.00	11.54	13.04	0.00	3.85	4.35	0.00
		6	19.23	21.74	0.00	11.54	13.04	0.00	11.54	8.70	33.33
		RS	15.38	17.39	0.00	15.38	13.04	33.33	7.69	4.35	33.33

Figure 1. Graphical illustration of the possible distortion of using raw score as an ability estimate.

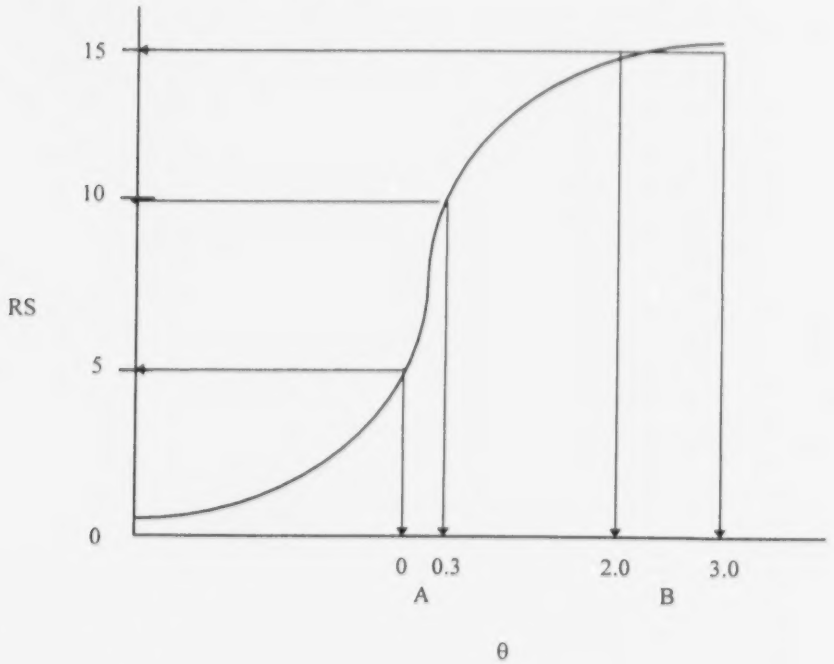


Figure 2a. Representation of the levels of comparsion for classifying score groups using total Raw Score (RS). The ovals represent each level of comparison on a 13 item test, a score-group for each possible raw score.

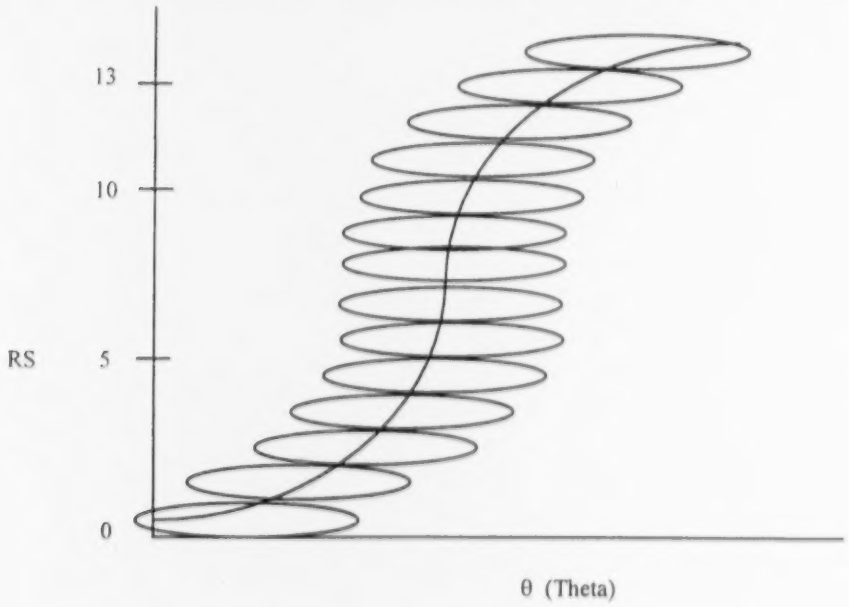


Figure 2b. Representation of the levels of comparsion using Theta for classifying groups for comparison. The rectangles represent each level of comparison on a 13 item test.

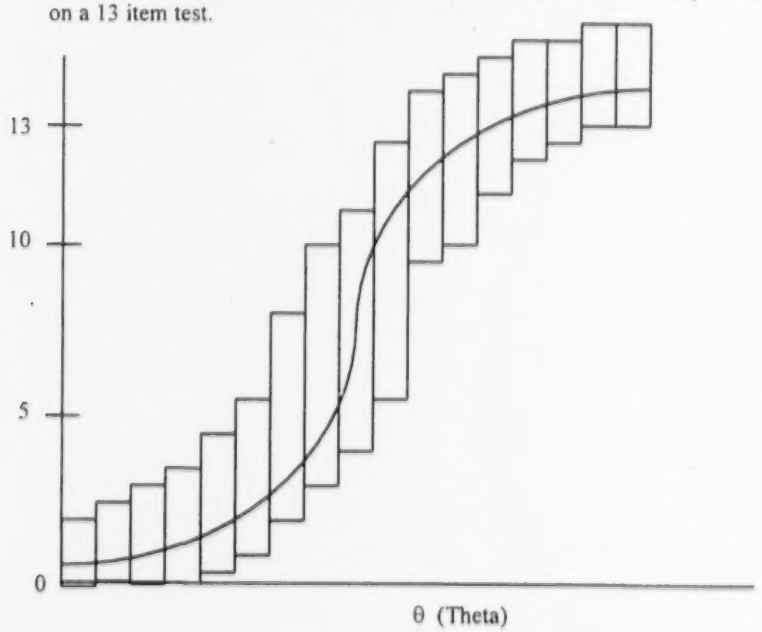


Figure 3. Graphical representation of the overall hit rates -- experimental condition.

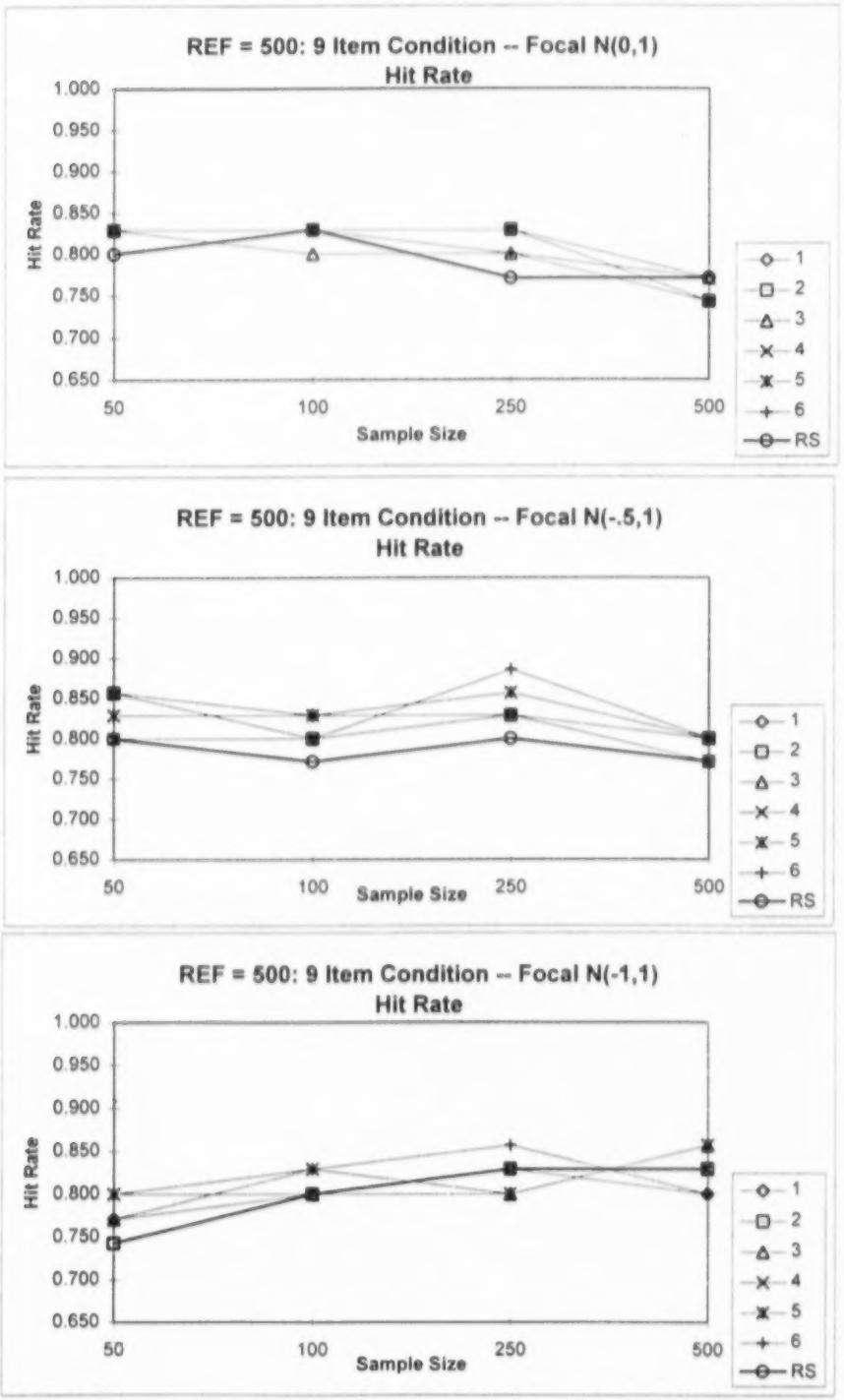
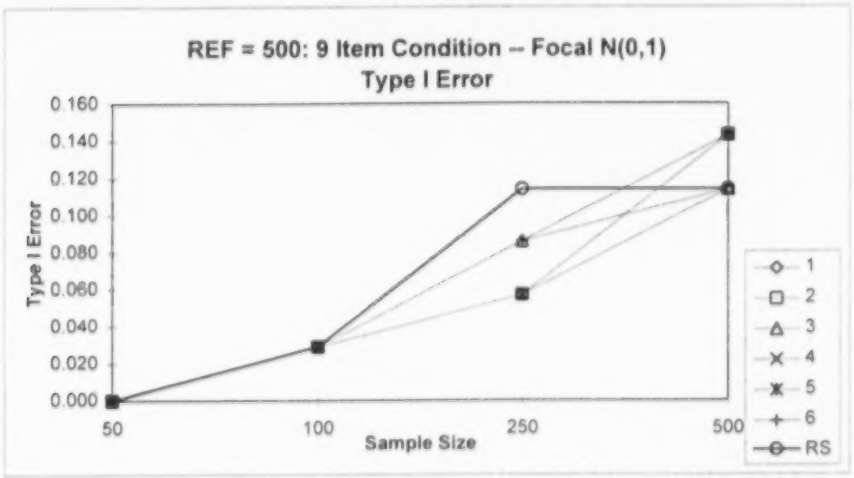
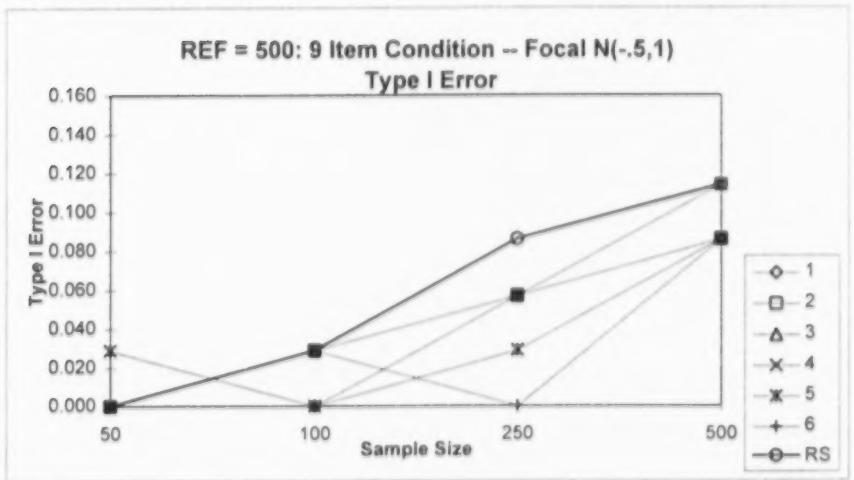


Figure 4. Graphical representation of the Type I error rates -- experimental condition.





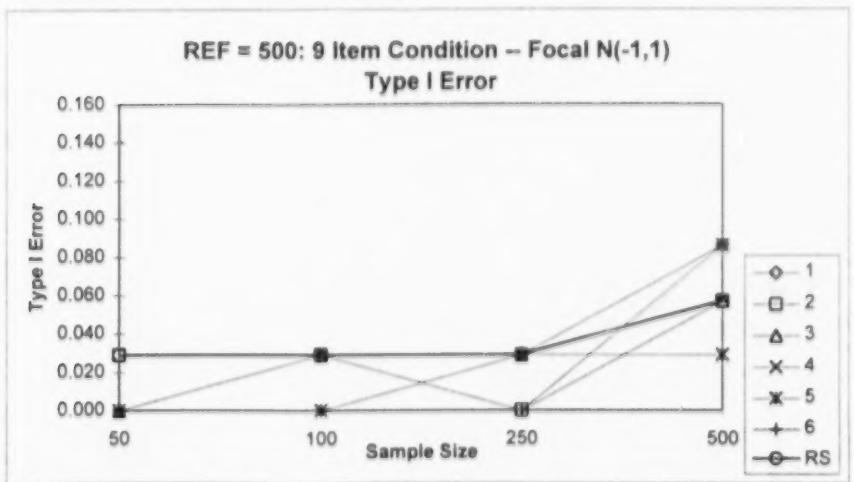
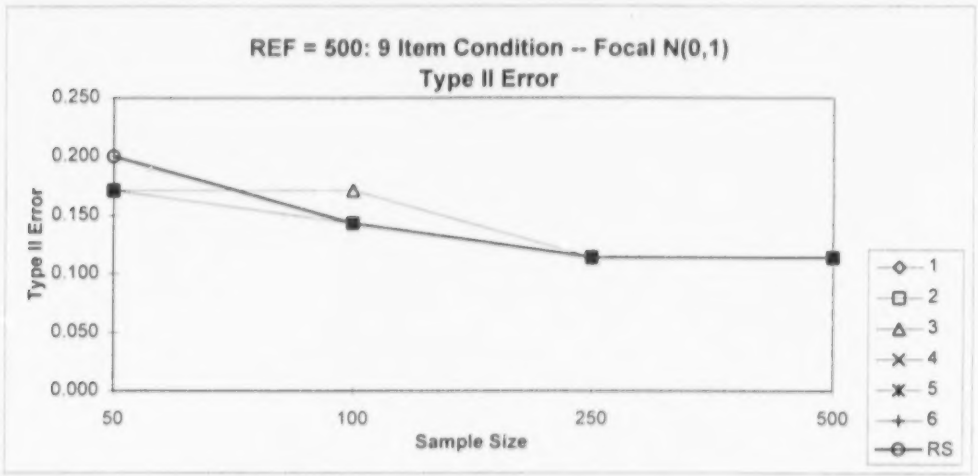
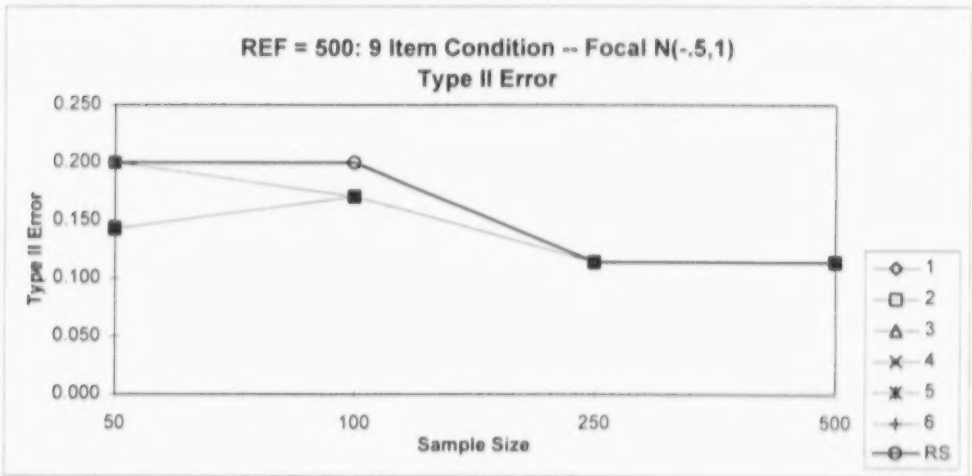


Figure 5. Graphical representation of the Type II error rates -- experimental condition.





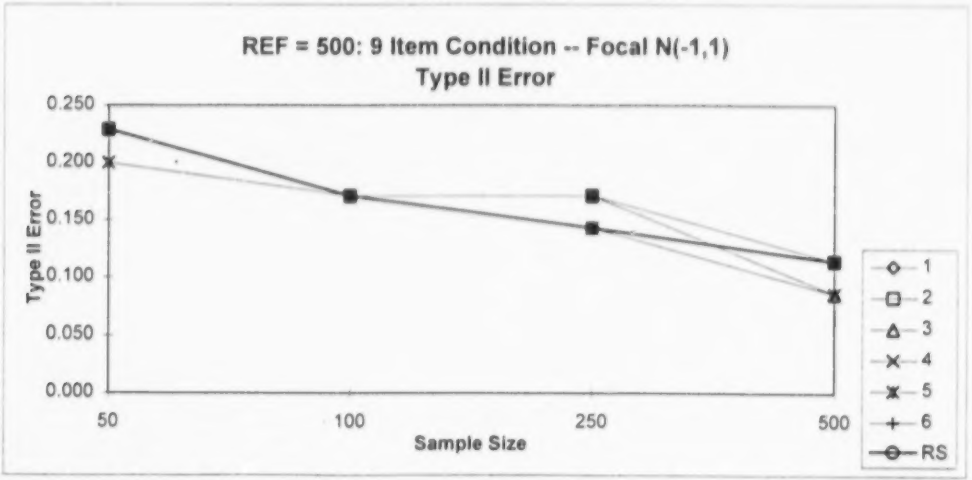


Figure 6. Graphical representation of the hit rates for adjusted items only --experimental condition.